Project title: Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and OWNer control
Project acronym: MOSAICrOWN
Funding scheme: H2020-ICT-2018-2
Topic: ICT-13-2018-2019
Project duration: January 2019 – December 2021

D3.1
First version of the reference metadata model

Editors: Aidan O Mahony (EISI)
Rigo Wenning (W3C)
Reviewers: Sabrina De Capitani di Vimercati (UNIMI)
Stefano Paraboschi (UNIBG)

Abstract

MOSAICrOWN assumes an enriched data market. In order to achieve this goal, the MOSAICrOWN system has to cover data and metadata where the metadata describe and give information about the data. This deliverable depicts a model high level architecture to ingest data and metadata. It also illustrates how to link data and metadata to inform the further processing of the data. Finally, this deliverable gives an initial list of metadata and vocabularies that the MOSAICrOWN use cases should take into account.

Type	Identifier	Dissemination	Date
Deliverable	D3.1	Public	2020.03.31

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825333.

MOSAICrOWN Consortium

1.	Università degli Studi di Milano	UNIMI	Italy
2.	EMC Information Systems International	EISI	Ireland
3.	Mastercard Europe	MC	Belgium
4.	SAP SE	SAP SE	Germany
5.	Università degli Studi di Bergamo	UNIBG	Italy
6.	GEIE ERCIM (Host of the W3C)	W3C	France

Versions

Version	Date	Description
0.1	2020.03.05	Document ready
0.2	2020.03.23	Revised document
0.3	2020.03.27	Second revised document
1.0	2020.03.31	Final version

List of Contributors

This document contains contributions from different MOSAICrOWN partners. Contributors for the chapters of this deliverable are presented in the following table.

Chapter	Author(s)
Executive Summary	Rigo Wenning (W3C), Aidan O Mahony (EISI)
Chapter 1: Introduction	Rigo Wenning (W3C), Aidan O Mahony (EISI)
Chapter 2: Data Ingestion	Daniel Bernau (SAP SE), Jonas Boehler (SAP SE), Aidan O Mahony (EISI), Michelle Mazzola(MC), Megan Wolf (MC)
Chapter 3: Data Governance in the Data Lake	Daniel Bernau (SAP SE), Jonas Boehler (SAP SE), Aidan O Mahony (EISI), Michelle Mazzola(MC), Megan Wolf (MC)
Chapter 4: Data Sharing and Analytics	Daniel Bernau (SAP SE), Jonas Boehler (SAP SE), Aidan O Mahony (EISI), Michelle Mazzola(MC), Megan Wolf (MC)
Chapter 5: Conclusions	Aidan O Mahony (EISI)

Contents

List of Figures

List of Tables

Executive Summary

One of the challenges facing MOSAICrOWN is how to describe the handling of the data being submitted to the enriched data market. How long should the data be stored in the data market? How should the data be represented? When were the data created? Metadata is one tool used in addressing this problem which also allows data owners to maintain control of their data.

MOSAICrOWN assumes an enriched data market that allows taking data protection into account by providing safeguards for personal data and valuable business data alike. This deliverable contains a high level thought model and architecture on how to ingest and process data and metadata in order to enable the data market to apply the relevant access rights, safeguards and use limitations to the data in the market. This deliverable leverages the requirements issued by WP2 and in turn WP2 receives the metadata model described in this deliverable which is applied to the use cases.

The metadata model described in this deliverable needs to facilitate three distinct phases. *Data Ingestion* is the process whereby the data owner submits their data to the MOSAICrOWN data market. *Data Governance in the data lake* is the concept of governing data with an aim of ensuring consistency, security, and availability as well as enforcing rights, standards, and policies as determined by data owners. Finally, *Data Sharing and Analytics* facilitates the data in the data market being made available for sharing and for analytics to be carried out within and outside the data market.

This deliverable constructs the metadata model by leveraging the use case requirements as described in D2.1 (“Requirements from the Use Cases”). Use Case 1 (“Protection of Sensitive Data in an Intelligent Connected Vehicle Data Market”) requires metadata to enable data wrapping and sanitization within the data market, ensuring consistency and quality of data, ensuring data are revoked in a timely fashion, and GDPR conformance. Use Case 2 (“Data Markets for Analysis of Financial Data”) requires financial related metadata specifying data wrapping, risk ratings, transaction related metadata, and traceability metadata. Use Case 3 (“Cloud-based Data Markets for Consumer Analytics”) requires sanitization metadata (e.g., method of sanitization), analytics-related metadata, and model-related metadata.

In order to achieve the goal of creating a metadata model, MOSAICrOWN has to cover data and metadata that tells more about the data. This deliverable depicts a high-level architecture model to help data handling with metadata. This includes ingesting data and metadata, how to link data and metadata to inform the further processing of the data and how to process the data to allow for sharing in the data market. Finally, this deliverable gives an initial list of metadata vocabularies that the MOSAICrOWN use cases should take into account. This includes privacy vocabularies, but also other useful metadata that help with the overall data handling. The list of vocabularies are an initial step, scoped by the use cases, to have the necessary metadata within the system and permits data about data quality, access control, privacy properties, licensing and more.

1. Introduction

The goal of MOSAICrOWN is to support the emergence of data markets leading to the development of a data economy. To do so, we need to identify key enablers in current markets and enable them in the digital world. Given our complex world, we will scope our efforts by our use cases. It is understood that a data market does not need to have all enablers present to function well. This in turn creates the need for modularity, combinability and extendability of the controlling tools such that the system follows the business imperatives negotiated between the partners of a given data market.

1.1 Purpose of this Deliverable

The purpose of this deliverable is to identify different vocabularies, concepts, and semantics needed for the use cases described in the requirements deliverable D2.1 ("Requirements from the Use Cases"). These will ultimately enable a data provider to control the data sharing within the market. The semantics provided by this deliverable will be instrumental in helping to fulfill the requirements from D2.1. As none of the known vocabularies can fulfill all the requirements, a combination of the vocabularies identified is required. MOSAICrOWN has different use cases to test the breadth of the options to create data markets and data value chains. This means that the metadata model has to work in diverse scenarios. Ideally, MOSAICrOWN could create interoperability even between its use cases.

1.2 Beyond access control

Controlling the access to and the flow of data requires more than metadata. It is not sufficient to know that Alice has access to information about Bob or that the maintenance company C has access to the data pool of the production line of company P. A protocol has to tell us how we proceed. This includes how and where to send 'login credentials' to, and how those are matched. Protocol work-flows are often integrated in existing languages like eXtensible Access Control Markup Language (XACML) [OAS09]. This means a metadata reference model not only needs to evaluate and integrate those languages, but also has to address the shortcomings and suggest alternative routes.

However, a data market requires more than basic access control to information. In particular, the ICV use case (UC1) will require even more than access control in the broad sense. The rules may not only touch on disclosure or non-disclosure of data items, but also may include rules about permitted uses and obligations, once the data have been shared or accessed. Known systems like Open Digital Rights Language (ODRL) [ISM$^+$17] have a certain expressiveness, but their semantics are rather tied to copyright and rights labeling in that field. The metadata model thus

has to explore the need for additional or even different semantics needed to have controlled data flows.

Finally, a data market can be implemented as a data hub that allows selective access and processing. Accessing such a data market will require the definition of an API: the functions and semantics the data market exposes to its users. But those single data markets are not the best way to implement data value chains. A central data market with access by its users will require an *always on* scenario. For the sharing of typical eCommerce information like shop profiles and credentials, this may work (e.g., using OAuth 2.0 [Har12] for the identification and XACML [OAS09] for the access control). In contrast, for an industry 4.0 scenario that has offline components, we have to envision another model where data are not only accessed on a hub, but extracted, packaged and sent through the data value chain. This also means that not only two parties are involved, but three or more actors who process, enrich and transform the information received. The question here is how to (i) package several data records together, (ii) add metadata information to the package, (iii) package metadata information together, (iv) secure and authenticate the package including its metadata, (v) and finally, ensure the appropriate expressiveness of the metadata given.

UC1 ("Protection of Sensitive Data in an Intelligent Connected Vehicle Data Market") is a complex use case. It will be the measure of the expressive power needed in order to share and exploit data in legal ways as there may be needs to identify context in order to cope with the rules stemming from regulation. The same could be true for UC2 ("Data Markets for Analysis of Financial Data") unless the data are duly anonymized, which gets more difficult over time with the constant improvement of database reconstruction attacks. UC3 ("Cloud-based Data Markets for Consumer Analytics") also needs less metadata as it mostly relies on anonymization. But if metadata are present during ingestion, both the financial and the consumer analytics cases can benefit from the semantics contained to apply adapted methods of anonymization. For example, calibrating the degree of perturbation in differential privacy or generalization in k-anonymity by privacy parameter ε or group size k.

This deliverable is a first approach to the metadata framework as details of the use cases mature. The intention is to create an overview of potential metadata vocabularies, how to ingest, link and process them in order to give the use cases the semantic expressions needed.

1.3 The metadata architecture

While not all use cases will follow a unique data flow, it was realized that reflecting on the data life cycle as expressed in Figure 1.1 helps greatly in the understanding of the processing steps involved. These steps include locating, ingesting and processing metadata, and eliciting necessary steps in determining how to handle the data before they enter the data market. Figure 1.1 is closer to a thought model than an architecture. It helps to identify steps for data enrichment and processing rules while data are on their way to the data market. This includes wrapping and sanitization techniques.

Figure 1.1 shows an approach with up to three major steps to arrive in the data market: data ingestion, processing in the data lake using data governance methods, and finally the metadata needed within the data market and the data governance therein. The mentioned thought model can inspire the following theoretic workflow. None of the use cases covers all of those steps, but those steps set the context to understand what the solutions for the specific use cases are:

1. Data ingested in the first step are linked to metadata which are also collected

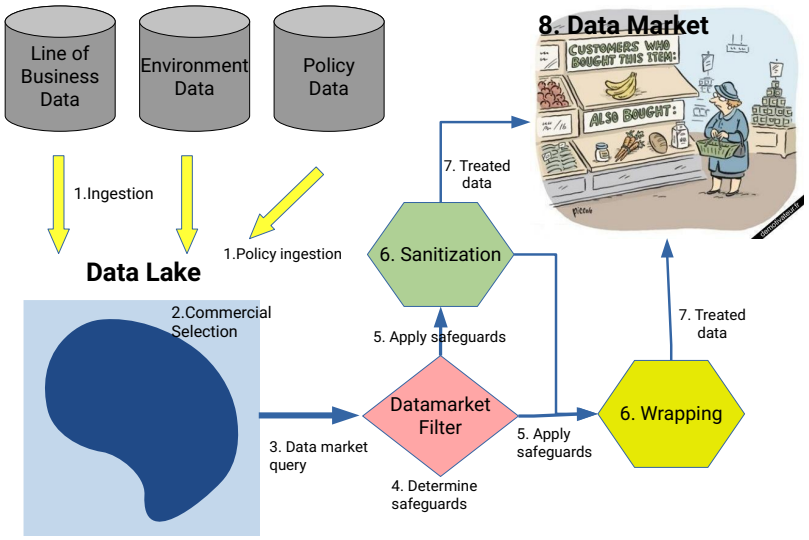

Figure 1.1: A flux towards the data market: A thought model

2. Data for the data market are selected by a function of commercial criteria
3. A query over the metadata linked to the data selected will disclose whether the data can be used for the intended purpose
4. In the case where the data cannot be used as is, additional safeguards are considered
5. Data are streamed into the necessary safeguarding process
6. Either data wrapping or data sanitization techniques are applied
7. The treated data will go into the data market
8. Even within the data market, queries can be informed and restricted by metadata delivered from the data lake to the data market.

Chapter 2 will talk about all the steps involved in getting data, metadata and policy data into the data lake. A data market is only as good as the data sources that feed into it, which means that data ingestion is decisive for both the data and the metadata. In UC2 and UC3 the data already exist in line of business applications or existing log files. Only in UC1 the data are directly collected from sensors. For this use case we must consider which sensors to include in the data collection. This means the data ingestion issues go beyond the transfer of data from line of business applications. Certain data points may not be useful and would only pollute the data lake. Consequently, Chapter 2 lists the data points and most of the metadata points that will be ingested into the data lake for further processing towards the data market in a second (logical) step.

Ingestion of data into the data lake takes all available instance data or available data categories into account. The step to ingest a data point into the data lake is a logical step, not necessarily a physical move of data. In fact, putting data into the data lake more or less means that the data have become managed data in the sense of MOSAICrOWN. The more data ingested into the data

lake, the more expressive power will be there to inform data management guidelines, sanitization techniques or wrapping techniques.

Using linked data [BHBL11] to attach metadata to data (illustrated in Figure 1.2), the data from various sources will be annotated by the metadata that relate to them. Thus the metadata available at collection or ingestion time augment the information about the data. A typical issue with current systems is that most of the policy information, like grounds for processing according to the GDPR, are lost later in the process. Not so here. The data and metadata will be ingested together with the managed data points. This concerns the full breadth of possible policy data including:

- Metadata about access control
- Metadata about policy restrictions, e.g., privacy
- Metadata about legal restrictions, like licenses
- Metadata about provenance and data quality
- Contextual data that are transformed into metadata, e.g., time and location

Further metadata found at collection time or within the line of business applications will be dependent on the use case. Eventually, the hope is to ingest enough metadata to create an enriched data lake. It remains undecided, and a matter of implementation, whether the data lake contains a knowledge graph or whether a less complex technology is used. The most important characteristic of this data lake in the thought model is that it can link data to their annotations. The annotations serve to link semantically computable properties to the data that inform later processing.

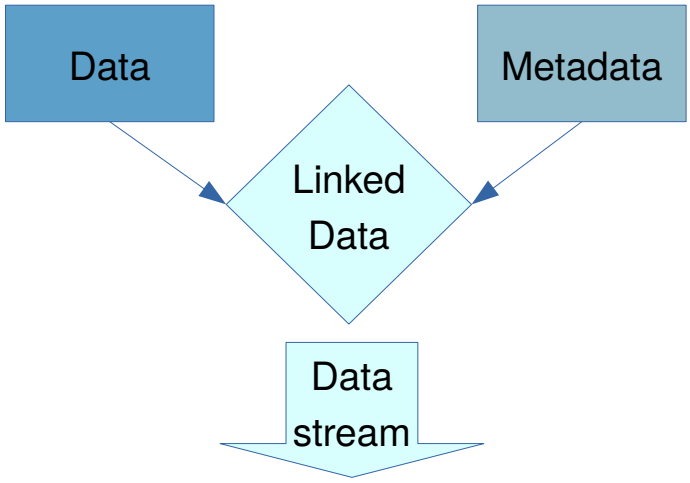

Figure 1.2: Using linked data to augment data with metadata

The data lake is a logical step in the process from data collection to data sharing. MOSAICrOWN has no fixed focus on a certain technology to achieve this. At the time of this deliverable, MOSAICrOWN intends to use linked data for some of the challenges, but it is not excluded

that a use case will do things differently. It is important to understand that the *data lake* is not necessarily a point of storage of all the data. For example, the SPECIAL project [KFD+18] implemented the filtering components as part of a data streaming setup. In the case of MOSAICrOWN, data will be altered or annotated on the fly before ending in the data market. There is no database in between.

With all data and metadata ingested, the data lake is used to select the data that go into the data market. In this second step, metadata queries can determine whether some data points or data category are suitable to be sent to the data market for further use. Categories, which will defined later for the use cases, can be designated either as data categories or metadata categories But the decision is not limited to sending or not sending data. In fact, the data lake is the (logical) place where data are rated according to their privacy sensitivity and other criteria. It is there that decisions are made to apply additional privacy safeguards, such as choosing whether data need to be encrypted or secured otherwise.

Once the data streams catering to the data lake are organized and augmented with metadata, the data can be processed in various ways. Depending on the use case and parts thereof, this ranges from simple queries to a complex result obtained from a reasoner (such as HermiT [GHM+14]) running over the data. Certain parts of the use cases can, on the other hand, be simpler (such as the enforcement of data at rest encryption). Requiring the use of a complex semantic system with a reasoner for those simple cases is not necessarily mandated. This deliverable argues for a pragmatic approach, where the most efficient solution should be selected, provided it can be expanded to interoperate with more complex parts of the data value chain.

Chapter 2 describes the categories, vocabularies and semantics helpful for the MOSAICrOWN use cases. It will also give hints on what to do in case new semantics are discovered or in case new semantics are needed for later transformations or data handling constraints.

Chapter 3 assumes that all relevant data have been ingested into the data lake. From there, new metadata relevant only at this stage need to be added, such as access control. This chapter also deals with the transforms that need to be done, before data can flow from the data lake to the data market as indicated in Figure 1.1.

Chapter 4 deals with the metadata that are created while using the data market. In fact, this does not only concern privacy, licensing and access control. It is also important to collect usage metadata and to feed this back into the data market and the data lake.

2. Acquiring Data into the Governance Framework

As already mentioned in Chapter 1, there are several ways to create a data market. While one way is to have a large data lake and manage access to its resources, the other way is to make data transportable and marketable by packaging them in certain ways. Both methods have in common that they require certain semantics, which means there is a need to organize the data flow and augment it with semantics. The semantics cater to algorithms that ensure compliance with legal requirements, the business model of the data market or the data value chain. MOSAICrOWN has three use cases and it has to be determined which of the models works best for each of them or whether they should all use the same model.

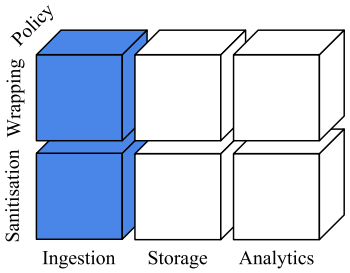

Figure 2.1: The MOSAICrOWN structure (ingestion)

In this chapter we consider the metadata required to satisfy the requirements with regard to the MOSAICrOWN data life-cycle stages highlighted in blue in Figure 2.1.

The three use cases of MOSAICrOWN have significant differences in their setup. While UC1 can already semantify data points at collection time, the other use cases collect information from traditional resources like line of business applications. At ingestion time from line of business applications there can be already some degree of processing to avoid ingestion of unwanted data, but also to apply the *semantic lifting* where needed and as described in Section 2.4.1.

The concept of a metadata structure helping with the management of data is only of reduced relevance for UC3 as described in D2.1. Metadata from the environment will not play a key role. But the concept of semantically augmenting data with metadata will still be useful as it will help to determine techniques and methods for an informed and (automatically or rule-based) controlled data sanitization.

The data markets for the analysis of UC2 data are in between the two other cases. The market is fueled by data from a variety of sources. However, those sources use traditional database technologies. A more intelligent data wrapping will have to analyze those traditional sources and add a step where the data streams from those sources into the data lake get semantified at some point.

All use cases have in common that they need access control. The semantification used to help with data wrapping may also be used to achieve a more fine granular level of access control taking

into account metadata and policy information.

In the following, thoughts are aligned with a proposed data life cycle. At every node of that data flow and life cycle, metadata play a role and have to be considered. The life cycle will assume a data lake that serves as a source for the data market. Data ingestion is there to provide fresh data to the data lake and keep it up to date. Once data are in the data lake, additional requirements from D2.1 such as access control and usage control can be taken into account. Finally, if the data are given to a downstream data controller[1], data need to be secured and the appropriate rights labeling needs to be in place for the downstream data controller to be able to behave lawfully.

Whether API or data packaging solution, it is important that during collection time, metadata are linked to the data to which they apply in order to later determine possibilities for re-use. It is possible that exceptions to this exist.

One exception may occur if only anonymized data are shared. In this case there are no privacy concerns anymore and GDPR does not apply [Eur16]. In principle, data that are safe against reconstruction attacks will not need to carry privacy semantics and other privacy use limitations. But MOSAICrOWN may still want to add usage rules and limitations in the form of metadata to the data for commercial reasons. Those will just have to pass a step where the sanitization and the anonymization is done. This step can be at any point before the output to the client of the data market.

In all other cases, there are many additional constraints which also affect data sharing. These constraints range from privacy over secrecy to storage or use limitations. If data are minted or collected from a source, some of the constraints just stem from the environment and should be preserved. Other constraints, especially those derived from the commercial exploitation and from business models, can be later added into the data lake, provided their connection to the data remains clear. Sometimes it can be sufficient to categorize data and attach the constraints to a given category. In other cases, this might not be of sufficient granularity and the policy elements have to be linked to the actual instance data concerned by them.

UC1 and UC2 have considerations about metadata ingestion. Requirements to that effect have been enumerated in D2.1, e.g., REQ-UC1-AC7: *Policies for data sets and platform users should be configurable by the data provider*. Ingestion of data can be done in many ways. For UC1, the sensors or some nodes between the sensors and the database are meant. For UC2 and UC3, the provenance of the data is less clear as such provenance was not necessarily recorded in the line of business application providing the data to the data lake. Data may come from multiple sources in multiple formats, which also raises interoperability issues at the edge of the data lake.

At the data ingestion time, the metadata ingestion constraints are mostly known. This corresponds to REQ-UC1-DI1: *MOSAICrOWN ingestion functions should support close to source deployment*. In an ideal world, the data source, e.g., for the electric car, would already inject the controlling semantics and the policy information into the system. But semantics and policy could also be injected at some node on the way into the managing platform.

All those requirements can be met either because the data source already provides the metadata in a format digestible to the data market or by adding that information from another source into the data lake before the transition to the data market.

[1] A downstream data controller is an entity who receives data from the data controller and processes them under its own control, but possibly with constraints set by the data controller [ABN+09].

2.1 Protection of Sensitive Data in an ICV Data Market (UC1)

The data handling approach via metadata is particularly useful in the ICV use case. An ICV will generate sensor data and environmental data and combines them with identity data, for example, by identifying the driver or the car or both. The metadata framework needs to identify:

- The protocols involved in sending data from the vehicle to the data lake
- The techniques that allow protocols to carry metadata
- Nodes where raw data streams can be combined with metadata streams
- The nodes for the *semantic lifting*, where URIs are added to everything that needs to be semantified. This includes, e.g., a UID for the vehicle in question.

We explored the work of the W3C Automotive WG, IG and Business Group to determine whether some of their terms and definitions could be re-used in our context. The W3C Automotive WG elaborates a vehicle data Specification [LAC+19] that serves the JavaScript interfaces for the car infotainment system. Those data points could be used as a foundation, but serialization and link between items for reasoning and metadata connection remain to be seen.

As vehicle data flows from the vehicle to the data lake, protocols need to be able to not only carry data, but also the corresponding metadata. This consideration is not relevant if the metadata are added from a different stream into the data lake. In the latter case, it will be a challenge to establish the link between the metadata and the data about which they express things. The ingestion process is illustrated in Figure 2.2.

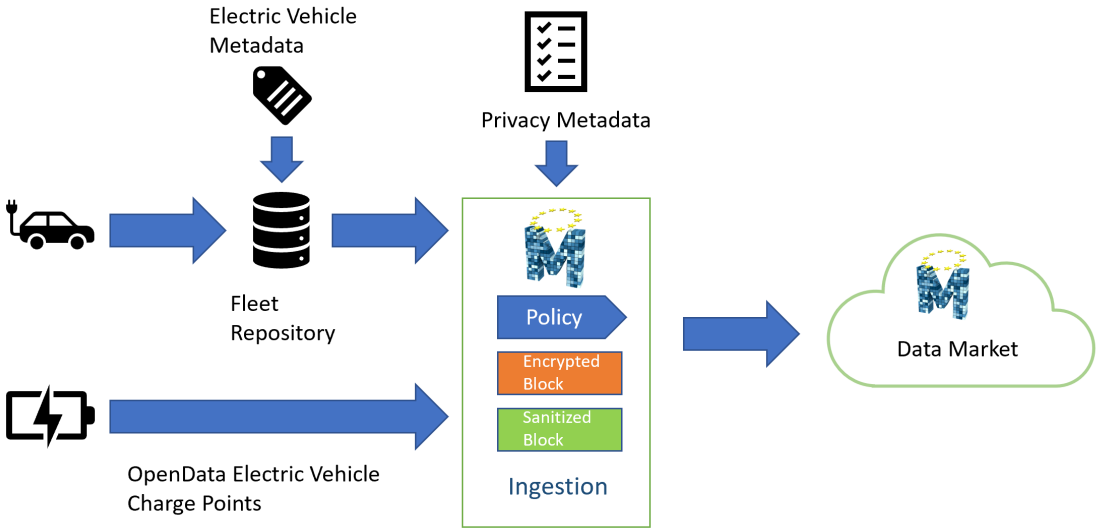

Figure 2.2: ICV data ingestion

As per D2.1 (“Requirements from the Use Cases”), there are two sources of data for ingestion in the ICV use case: the fleet data repository and the open data repository containing Electric Vehicle (EV) charging point locations and parking data. For the vehicle data points, a certain amount of vehicle specific metadata are required to be applied by either the fleet data repository or by the vehicle itself prior to submission to the fleet data repository. The ICV metadata are based on existing vocabularies and the categories are introduced in Table 2.1.

Category	Standards Body	Description
Vehicle	FIWARE	Vehicle data model [Pro19b]
GDPR	W3C	DPVCG GDPR Legal Basis Vocabulary [DPV19]
Privacy	W3C	Data Privacy Vocabulary [Dat19]
PROV	W3C	Data Provenance Vocabulary [W3C13]
WebCryptoAPI	W3C	Encryption Related Vocabulary [Wat17]
Quality	W3C	Data Quality Vocabulary [AI16]

Table 2.1: Metadata vocabularies

2.1.1 Electric vehicle data points

The EV data points originate from the vehicle itself and are accessed via an API. The data points required to satisfy the requirement of D2.1 [BW19] are detailed in Table 2.2 and are based on data points detailed in the FIWARE vehicle model [Veh20], and the FIWARE EV charging point model [Pro19a].

Category	Vehicle Data	Description
Fundamentals	Vehicle Status	Basic status information
	Capabilities	Vehicle capabilities
	Failure Message	Failure reason
	Firmware Message	Firmware version
	Historical State	Historical State
	Vehicle Status	Current status
Chassis	Charging State	Range, battery level, etc
Diagnostics	Diagnostic State	Speed, battery voltage, etc
	Maintenance	Battery service call date, etc
	Usage	Last trip battery remaining, etc
Infrastructure	Home Charger State	Charging power, solar charging, etc
Points of Interest	Navi Destination	Navigation destination
	Vehicle Location	Current vehicle information
	Vehicle Time	Current vehicle time

Table 2.2: Electric vehicle data points

2.1.2 Electric vehicle metadata

A certain amount of metadata needs to be added to the data points read from the EV. These metadata are based on the FIWARE vehicle data model [SDM19] and are illustrated in Table 2.3. These metadata augment the ICV data at ingestion.

Title	Description
UID	Unique identifier
Type	Entity type. It must be equal to Vehicle
Source	A sequence of characters giving the source of the entity data
Name	Name given to this vehicle
Description	Vehicle description
Vehicle Type	Type of vehicle from the point of view of its structural characteristics
Category	Vehicle category(ies) from an external point of view
Location	Vehicle's last known location represented by a GeoJSON Point
Previous Location	Vehicle's previous location represented by a GeoJSON Point
Heading	Direction of travel of the vehicle
VIN	Vehicle Identification Number (VIN)
Vehicle Plate Identifier	An identifier or code displayed on a vehicle registration
Fleet Vehicle Id	Identifier of the vehicle in the context of the fleet of vehicles to which it belongs
Owner	Vehicle's owner
Date Modified	Last update timestamp of this entity
Date Created	Creation timestamp of this entity

Table 2.3: Electric vehicle metadata

2.1.3 Electric vehicle charging point data

The EV charging point data are taken from Ireland's Open Data Portal [Cou18] and provide the details of electric charging points in Cork City in the Republic of Ireland. There are two types of chargers in the city: Standard chargers and Fast Charge charging points. The locations are provided in longitude and latitude. Further details as to the structure of these data are provided in Table 2.4.

Title	Description
ObjectID	ID of Charging Point
Name	Name of Charging Point
Details	Detailed information on charging point
Longitude	Longitude
Latitude	Latitude
Type	Fast charge/standard etc

Table 2.4: OpenData electric vehicle charge points

2.1.4 Privacy related metadata for ICV use case

The categories and metadata are based on [Dat19] and [DPV19] and relate to the driver of the ICV. The entries in Table 2.5 are not exhaustive as some data may be derived from analytics. Also, in Table 2.6, the GDPR related vocabulary is presented.

Category	Class	Description
Personal Data	Country	Information about an individuals country
	Location	Information about an individuals location
	MACAddress	A category of personal data
	UID	A unique identifier used to identify an individual
	IPAddress	An Internet protocol address of a device used by an individual
	Age	Information about an individual's age
	Biometric	Biometric information on an individual
	Contact	Information that can be used for contacting an individual
	Credit Card	Information about an individual's credit card number
	Email Address	Information about an individual's Email address
Purposes	Context	Used to scope the purpose
	Sell Data To Third Parties	To sell data or information to third parties
	Sell Insights From Data	To sell or commercially provide insights obtained from analysis of data

Table 2.5: Data privacy metadata

Class	Description
A6-1-a-explicit-consent	Explicit Consent
A6-1-a-non-explicit-consent	Regular Consent, consent but not at the level of being 'explicit'

Table 2.6: GDPR legal basis metadata

The EV data points combined with the EV metadata constitute the EV data which will be ingested into the data lake. The EV data are furthermore annotated with Privacy Metadata at the ingestion stage. The following example shows a subset of EV metadata in JSON and then shows the example of EV privacy metadata indicating the data privacy permission to sell data or information to third parties of the personally identifiable data.

```
Content-Type: application/ld+json
{
    "ev-meta-data": {
        "uid": 169619866884,
        "veh_name": "testjlr",
        "veh_description": "2019 4 door Sample JLR EV",
        "veh_vin": "1HMA17A28H3896213",
        "veh_plate_id": "B-HM-7836",
        "veh_location": {
            "latitude": 48.7358724567964,
            "longitude": 9.04108692828577
        },
        "timestamps": {
            "creation": {
                "unix_format": 1584103820.0,
                "reg": "2020-03-13 12:50:20.286058"
            }
        }
    }

    "ev-dpv-meta-data": {
        "veh_name": {
            "dpv:SellDataToThirdParties" : "true"
        }
    }
}
```

Data provenance is a consideration for ICV data (and data in general) from the perspective of both data generated as part of the MOSAICrOWN data analytics as well as data analytics conducted on sanitized data. The vocabulary which will be used in the data market is presented in Table 2.7.

Component	Type	Description
Entities	Entity	Physical, digital, conceptual, or other kind of thing with some fixed aspects
Activities	Activity	Something that occurs over a period of time and acts upon or with entities
	Generation	Completion of production of a new entity by an activity
	Usage	Beginning of utilizing an entity by an activity
	Communication	Exchange of some unspecified entity by two activities
	Start	When an activity is deemed to have been started by an entity, known as trigger
	End	When an activity is deemed to have been ended by an entity, known as trigger
	Invalidation	Start of the destruction, cessation, or expiry of an existing entity by an activity

Derivations	Derivation	Transformation of an entity into another
	Revision	Derivation for which the resulting entity is a revised version of some original
	Quotation	The repeat of (some or all of) an entity, such as text or image, by someone who may or may not be its original author
	Primary Source	Produced by some agent with direct experience and knowledge about the topic
Agents, Responsibility, Influence	Agent	Something that bears some form of responsibility for an activity taking place
	Attribution	Ascribing of an entity to an agent
	Association	Assignment of responsibility to an agent for an activity
	Delegation	Assignment of authority and responsibility to an agent
	Plan	An entity that represents a set of actions or steps intended by one or more agents to achieve some goals
	Person	Agents are also of type Person
	Organization	Organization is a social or legal institution such as a company, society, etc.
	Influence	Capacity of an entity, activity, or agent to have an effect
Alternate	Alternate	Two alternate entities present aspects of the same thing
	Specialization	Shares all aspects of the latter, and additionally presents more specific aspects of the same thing as the latter
Collections	Collection	Entity that provides a structure to some constituents that must themselves be entities
	Membership	Stating the members of a Collection

Table 2.7: Data provenance metadata

At ingestion time, the W3C WebCryptoAPI [Wat17] provides a vocabulary for describing the cryptographic algorithms required to secure data both at rest and in flight. The vocabulary specifies a list of supported operations (e.g., encrypt, decrypt, sign, etc).

The impact of sanitization on data quality is well documented. It is desirable to be able to specify the metrics and measurements related to data quality and a vocabulary covering quality related metadata is presented in Table 2.8.

Class	**Description**
Quality Measurement	Represents the evaluation of a given dataset (or dataset distribution) against a specific quality metric
Metric	Represents a standard to measure a quality dimension
Dimension	Criteria relevant for assessing quality
Category	A group of quality dimensions in which a common type of information is used as quality indicator

Quality Policy	Policy or agreement relating properties and classes of policy-dedicated vocabularies, such as ODRL[ISM$^+$17]
Quality Annotation	Quality annotations, including ratings, quality certificates or feedback that can be associated to datasets or distributions
Quality Metadata	Defined to group quality certificates, policies, measurements and annotations under a named graph

Table 2.8: Quality related metadata

2.2 Data markets for analysis of financial data (UC2)

The financial data market needs to define data that are sensitive from a privacy standpoint or otherwise not wanted for the ingestion into the data lake. When ingesting data from different sources, REQ-UC2-S1 requires that the system knows how to avoid ingesting personal or sensitive data. This can only be done if the system knows what data are sensitive or personal. In order to match, the system needs to define:

- What are sensitive data in a taxonomy or ontology. As the financial industry is a very regulated one, there are concrete guidelines on what constitutes sensitive data. The natural language lists will have to be transformed into an ontology.
- What are personal data in a taxonomy. MOSAICrOWN can use the Data protection vocabulary developed by the Data Privacy Vocabulary Community Group in W3C [PP19].

Sensitive data are legally defined by Art. 9 of the GDPR and encompasses racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data. But the goal in UC2 is to also capture information that may not be sensitive in the sense of the GDPR, but is sensitive in the more common language sense as creating a risk of damage or embarrassment. For UC2, those terms are known in the financial world, sometimes called “exposed persons”, but there is no open public ontology for those terms. The classification rather seems to be an asset and know-how by every financial actor.

Consequently, UC2 consists in matching the families of semantics to each other. On the one hand there is the data protection sensitivity that can be derived from the semantics of the Data Privacy Vocabulary [PP19]. The vocabulary does not express the amount of risk itself as this may even be context dependent. However, it will be easy to add a risk rating in the scale 1 – 10 to the categories involved in the processing. Note that this can happen at the time at which data are acquired into the data lake or later. On the other hand, depending on the data to be shared in the analytics environment, financial semantics and data categories will have to be augmented by the data protection semantics.

Data categories will have to be matched with financial semantics for the data used in UC2 and coming from the line of business applications. Integration of provenance information [W3C13] will ensure the traceability of the data sources and allow to make quality and risk assessments.

Finally, the semantics of the financial data have to be assigned during the ingestion from the line of business application. A financial ontology is needed. There are several financial ontologies around. The relevant ontologies include:

- The OMG Financial Business Ontology (FIBO) [Fin17]
- ISO 20022 [BKdM12], especially the ontology on card transactions
- Semantics that can be drawn from XBRL, the eXtensible Business Reporting Language [Liu13]
- Semantics drawn from the Financial Products Markup Language (FpML) [JG13]

For the anonymization at ingestion time, the financial industry use case uses the same techniques as the one on consumer analytics detailed in Section 2.3.

It should be made clear that this section outlines a suggested framework that will ultimately vary based on the client and situational needs around each, individual use case. While defining general ontologies and setting a potential framework are critical for the overall progression of this project, it should be clear that this can be added on to, or altered dependent upon the demonstrated need of the use case.

Sample financial data are presented in Table 2.9.

Category	**Class**	**Description**
Transaction	Date	The date of the transaction
	Time	The time of the transaction
	Merchant	The merchant at which the transaction occurred
	Amount	The monetary amount
	Currency	The currency in which the monetary amount is expressed
Cardholder	Geography	The geographic area the cardholder is living in
	Demographic	The age/gender/race/etc. of the cardholder
	Attributes	Various attributes specific to the cardholder and transaction (cardholder presence, cross border, point of interaction, etc.)
Card/Account	Open Date	The date the card was opened
	Credit Limit	The credit limit of the card issued
	Product Type	The type of card that was issued (Platinum Credit, Standard Debit, etc.)
	Balance	The balance of the account
	Card Number	The card identifier (includes full number, BIN, any card identifier)
	Issuing Bank	The bank that issued the card (including any additional information on that bank)

Table 2.9: Transaction related metadata

While categorization of the financial data is important, policy metadata and processes are critical for UC2. The policy metadata describes the steps necessary to meet policy and privacy standards set by the company, global governments and financial regulators. Below, there is an example for a specific field name, PAN, that can serve as a template for other field names. It should be noted that both "compliance type" and "data usage" are broader items and would result in further output and action to ensure the data field does, in fact, meet the policies in place.

Altogether, the metadata should contain the following information:

- Data owner - owner of the transaction level data
- Data category - data category for the specific set (Clearing, Auth, etc.)
- Field name - title of the field specified in the metadata
- Compliance categorization - category for which the data are compliant and the corresponding actions necessary to get to this level of compliance (Payment Card Industry (PCI), GDPR, etc.)
- Sensitivity level - metric to measure the overall sensitivity level of the data, on a range from 0 (non-sensitive) to 1 (highly sensitive)
- Data usage - what the data can be used for and specific instructions regarding protecting non-provisioned usage (i.e. broad usage, specific platform usage, admin only usage, no usage, etc.)

The following example shows how this could be implemented for a given data category during ingestion.

```
{
    "data owner": "admin",
    "data category": "clearing",
    "field name": "PAN",
    "compliance categorization": "<complianceType>",
    "sensitivity level": 0.9,
    "data usage": "<dataUsage>"
}
```

2.3 Cloud-based data markets for consumer analytics (UC3)

The basic idea behind UC3 is that businesses need to share data including trends or customer needs in order to evaluate a possible cooperation or to serve as an argument in business negotiations. In order to avoid the sharing of sensitive business and personal data, data are anonymized either before entering the data market or before releasing any aggregate statistics computed from the data in the data market to enable cooperation of companies over non-personal, anonymized data.

With ever more sophisticated reconstruction attacks and linkage attacks using publicly available data, the challenge for strong anonymization with meaningful protection guarantees is ever increasing. Thus, support for choosing and interpreting privacy parameters is required. Furthermore, complementary deployments of cryptographic protocols and anonymization means need to

be considered to not only protect the inference that can be made from data, but also the secrecy of the data themselves.

This use case needs metadata in order to describe

- Confidentiality requirements
- Sanitization requirements
- Semantics about access control
- Semantics about use and processing limitations

Confidentiality requirements

Confidentiality metadata might describe cryptographic protocols used to provide collaborative statistics in a distributed setting, e.g., two companies want to securely compute an anonymized statistical value (enterprise benchmarking).

Sanitization requirements

Sanitization metadata mainly describe what kind of anonymization method is applied on which parts of the data. Additionally, the configuration for the method allows specifying parameters for the anonymization method (e.g., ε for differential privacy). As this configuration requires some expert knowledge, we will provide abstractions, in the form of easier to understand privacy guarantees (or re-identification risks) and preset values based on parameters commonly used in the scientific literature. Also, for some processing the data types per column (e.g., String, Double) might be required if the anonymized data should remain the same data type as the input data.

Altogether, the metadata should contain the following information:

- `method`: the anonymization method (e.g., perturbation or probabilistic selection via differential privacy)
- `dataSrc`: the identifier of the dataset that should be anonymized
- `columnTypes`: optionally, the data types of the columns in the dataset
- `methodConfig`: the configuration of the anonymization method
  - `configParams`: parameters for the anonymization method, e.g., the privacy parameter ε for differential privacy or an alternative description of the privacy guarantee (or bound on the potential privacy loss)
  - `targetColumn`: the column on which the anonymization should be applied
  - `sensitivity`: measure of how accurately queries results can be published while preserving a desired level of privacy

The following example shows metadata in JSON to anonymize a data set with two columns, the first containing a String, the second a Double value and only the second column requires anonymization (note that column indexing begins with 0):

```
{
  "method" : "differentialPrivacy",
  "dataSrc" : "<dataSetId>",
  "columnTypes" : "String,Double",
  "methodConfig" : {
    "epsilon" : 0.1,
    "sensitivity" : 1,
    "targetColumn" : 1
  }
}
```

Access control

Additional metadata for access control might describe *who* (subject) can access *what* (resource), and *how* (privilege):

- `subject`: subject requesting access
- `resource`: resource for which access is requested
- `privilege`: restrictions on the access, e.g., only anonymized data with certain protections (see sanitization requirements)

Use and processing limitations

From a technical perspective UC3 distinguishes between means for *microdata* and *macrodata* anonymization. The former receives as input n records and outputs n anonymized records, whereas the latter receives n input records and outputs $m \neq n$ anonymized records (mainly $m = 1$). An example application for microdata anonymization is the sanitization of a salary database. Here, each record is perturbed by adding random noise to hide individual values, but allow statistical aggregations such as computing the mean. In contrast, for macrodata anonymization an original dataset is used as input for an aggregation function (as found in machine learning), and the output (or an intermediate representation) is anonymized. An example for macrodata anonymization is the computation of a median value that is perturbed with differential privacy (or an anonymized machine learning model).

From the legal perspective according to GDPR Art. 5 1(e) (emphasis ours): “personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, *scientific or historical research purposes* or *statistical purposes* in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject (‘storage limitation’)”. This requirement for having a justified purpose for personal data especially affects macrodata anonymization since the data have to be gathered before being anonymized, and thus the original data might be stored in the data marketplace over a limited time frame before the anonymized macrodata is released.

All in all, the two mentioned processing approaches for anonymization, and the legal purpose requirement lead to the following metadata requirements:

- `processingType`: e.g., microdata, macrodata
- `purpose`: e.g., scientific, or statistical use

2.4 Ingestion API

From the high-level model in Section 1.3 we derive that there is a need to provide some technical way in order to transfer the data from a line of business application into the data lake. Because the line of business application does not contain the relevant policy information that allows the system to draw conclusions on what to process, data from the line of business application have to be augmented with metadata and policy information when entering the data lake.

2.4.1 Semantification

The Big Data Europe Project [ASV+17] talks about *semantification* of the data. Non-semantic data will mainly come from line of business applications. According to Mami et al. [MSAV16], the semantic lifting of non-semantic data can be achieved through the integration of mapping techniques e.g., R2RM3 [W3C12], CSVW4 annotation models [PTKH15] or JSON-LD contexts [SLK+19]. This integration can lead to either a representation of the non-semantic data in RDF, or its annotation with semantic mappings so as to enable full conversion at a later stage.

Within the data lake, it is important to also follow the recommendations from [MSAV16] to preserve the semantic enrichment as much as possible. RDF-based data representations and mappings have the advantage (e.g., compared to XML) of using fine-grained formalisms (e.g., RDF triples or R2RML triple maps) that persist even when the data themselves are significantly altered or aggregated. They also suggest best practices for the semantification using the IRI's [DS05] and literals.

Especially for UC2, MOSAICrOWN may benefit from the results of the OpenBudget project[2] which already has a large number of recipes to semantify financial data.

It is preferred to transform the data to RDF as the data handling tool chain for complex scenarios will be using Linked data (and thus RDF), but this is not a requirement or necessary condition for the system to function. If there is no transformation to RDF, the annotations done to the original data with mappings should be done in a way that still allows a later transposition into the RDF format. Especially, the data lake needs to preserve the semantics and the relation between data and metadata. Once put in the data market, data do not have to be in the same format as can be seen in Chapter 4.

2.4.2 Policy ingestion

The data lake also serves to augment data with policy data. This policy data can come directly from the data sources that are ingested, or can be added later on. If the existing system feeding into the data market already contains policy information, this has to be ingested with the data. Also, it may be augmented further in later processing steps. Deliverable D3.3 ("First version of policy specification language and model") will describe the first version of the policy language that will be used in MOSAICrOWN.

2.4.3 Anonymization, sanitization and wrapping

Especially for UC2, but also for UC3, the added metadata will have to provide semantics to distinguish between highly-sensitive (personal) data and less sensitive data. For highly-sensitive (personal) data, a more advanced sanitization may be required. As suggested in Section 2.2, an

[2] `http://openbudgets.eu/about/deliverables/`

evaluation of sensitivity of data can already be made at data ingestion time. Such an evaluation will be usually less sophisticated than those being done in the data lake created.

Depending on the use case, it may be desirable not to ingest the full breadth of a line of business application into the data lake, but apply measures during the transfer such as:

- Excluding the transfer of certain categories of data that carry a higher risk
- Apply anonymization techniques before ingesting data into the lake
- Apply wrapping techniques to secure data in the lake

These measures do not exclude later treatment when transferring data from the data lake into the data market. The sanitization or wrapping at ingestion time described above is rather a pre-computing of data that is selected for exploitation. This pre-computing can be part of a batch computation step within the system.

3. Data Governance in the Data Lake

Chapter 2 investigated the metadata and policy information to add to the data lake. This chapter will investigate the data transformation and the metadata needed within the data lake once the data is ingested. Like in a funnel, the closer we get to the data market, the smaller the amount of metadata addition will be, because most of the needed information was already ingested. As the data converges with the data market, the more the system's focus changes from information collection towards helping to install a data market that is informed by the constraints stemming from the ingested metadata. Therefore, this chapter will look into what semantics are required to manage data access and data processing in the data lake. This may lead to transform affecting data, but especially data that are later transferred to the data market.

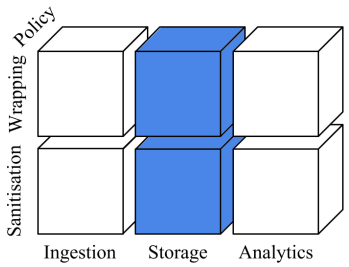

Figure 3.1: The MOSAICrOWN structure (data lake)

In this chapter we consider the metadata required to satisfy the requirements with regard to the MOSAICrOWN data life-cycle stages highlighted in blue in Figure 3.1.

3.1 Threat modeling in the data lake

Metadata in the data lake should help the systematic analysis of threats against assets. Assets can be privacy, data protection, but also commercial assets like intellectual property, sui generis rights on data, etc. Once some relation between the presence of certain data categories or certain metadata and a threat is identified, the same mitigation may be applied automatically in the future via an integration of that relation into the Linked Data relations. This allows the treatment of big data without the need for constant human intervention. MOSAICrOWN is concentrated on threats to privacy and those will be the ones being researched.

At the end of a threat analysis (such as the one illustrated in Figure 3.2 [WJ15]), a mitigation strategy will be identified and recommendations will be made.

3.2 Access control

Access control in a narrow sense is a very basic concept of data governance because it is binary. However, over time, it has become complex as human social organization with regards access

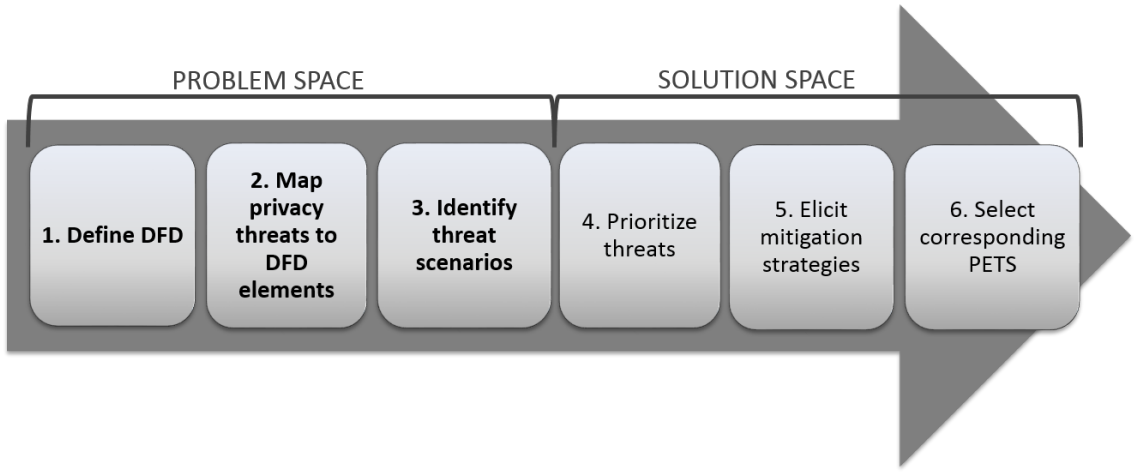

Figure 3.2: The LINDDUN methodology steps

follows a complex set up of those rules. In UC2 and UC3, access control also acts as client management via access credentials. But here, we certainly need a basic access control management for the data lake in the thought model.

In the data lake, this means that data should only be accessible after the identity of the data requester is confirmed. Authentication is the process by which the identity of the requester is confirmed. Authorization confirms that the requester has the right to execute a service.

The most widespread industry standards in this area include:

- The eXtensible Access Control Language XACML Specification [Mos05]
- Oauth2 [Har12][1]
- SAML 2.0 [CKPM05]

3.3 Usage control

Usage control is a well understood and well known technique in the context of EU projects. It started with the PRIME project EU - Project 2004, was also used in MIT's TAMI-Project[2], was further adapted to the web services world with the PrimeLife EU Project[3] and, was updated and further evolved for Big data by the SPECIAL project[4].

But the usage control in MOSAICrOWN goes beyond privacy. There is also a need to express further constraints for the processing of data, especially in the area of licensing. To express those, MOSAICrOWN will consider the adoption of the Open Rights Description Language ODRL [ISM+17].

[1] https://oauth.net/2/
[2] https://www.w3.org/2005/01/TAMI-prospectus.html
[3] http://primelife.ercim.eu/
[4] https://specialprivacy.eu/

3.4 Data governance metadata for ICV use case (UC1)

Vehicle data will go from the sensors to the data lake and from the data lake to the data market. Within the data lake, data privacy metadata will be added to the payload data. The use case will use the Data Privacy Vocabulary and the MOSAICrOWN policy language. The categories and metadata were based on the W3C Data Privacy Vocabulary [Dat19] and are shown in Table 3.1.

Category	Metadata	Description
Base	Data Controller	Data Controllers that control this particular data handling
	Data Subject	Class of Data Subject that this particular data handling applies to
	Personal Data Category	A category of personal data (as defined by GDPR article 4.1) from the personal data categories taxonomy
	Personal Data Handling	Top Class to describe a concrete instance of legal personal Data Handling
	Processing	A type of processing from one of the processing categories
Processing	Adapt	To modify the data, often rewritten into a new form for a new use
	Alter	To change the data in a small but significant way
	Anonymize	Irreversibly alter personal data
	Combine	To join or merge data
	Consult	To consult or query data
	Copy	To produce an exact reproduction of the data
	Derive	To create new derivative data from the original data
	Destruct	To process data in a way they no longer exist
	Disseminate	To spread data throughout
	Erase	To delete data
	Move	To move data from one location to another including deleting the original copy
	Profiling	Create a profile that describes or represents a person
	Pseudo anonymize	Replace personal identifiable information by artificial identifiers
	Remove	To destruct or erase data
	Restrict	To apply a restriction on the processing of specific records
	Retrieve	To retrieve data, often in an automated manner
	Share	To give data (or a portion of them) to others
	Store	To keep data for future use
	Structure	To arrange data according to a structure
	Transfer	To move data from one place to another
	Transform	To change the form or nature of data
	Transmit	To send out data

	Use	To use data
Technical and Organizational Measures	Anonymization	Process by which personal data are irreversibly altered
	Authentication Protocols	Protocols involving validation of identity
	De Identification	Process by identifiable personal data (PII) are converted to un-identifiable personal data
	Encryption In Rest	Encryption of data when being stored
	Encryption In Transfer	Encryption of data in transit
	NDA	Non-disclosure Agreements
	Storage Deletion	Defines how secure deletion is guaranteed
	Storage Duration	A duration denoting limitation on storage of personal data
	Storage Location	Defines a location or geospatial scope, where the data are (physically) stored
	Storage Restoration	Regularity and temporal span of data restoration/backup mechanisms that guarantee that data are preserved
	Storage Restriction	Restrictions required or followed regarding storage of data

Table 3.1: Data storage metadata for UC1

An example of some of these terms is below. This example shows metadata relevant to storage based off of UC1 requirements.

- Destruct - Permission to destroy the data being referenced
- Transfer - Permission to transfer the data being referenced
- Encryption In Rest - Method for encrypting data in rest

```
"ev-dpv-meta-data": {
    "veh_name": {
        "dpv:Destruct" : "true",
        "dpv:Transfer" : "true",
        "dpv:EncryptionInRest" : "AES-256-CBC",
    }
}
```

3.5 Data governance metadata for the financial use case (UC2)

The MOSAICrOWN financial use case has the advantage that its data are stored in a structured way in line of business applications. The ingestion into the data lake provided additional metadata and policy information. In the area of the data lake and for data governance, it is the question of finding categories of data within the line of business applications that are suitable for the data market and

determine how to transform the data to make them marketable. MOSAICrOWN provides for sanitization, anonymization and wrapping techniques to implement those transformations.

All data from the line of business application are thus matched against the W3C Data Privacy Vocabulary [Dat19]. Each category will then find a privacy threat rating attached to it. An ontology using the MOSAICrOWN policy language (D3.3 "First version of policy specification language and model") will then link the categories to specific techniques in order to trigger the appropriate transform.

Category	Metadata	Description
Data Wrapping Base	Sanitize	The process of cleaning the data
	Classify	The process of classifying data based on security necessary
	Storage	Where data will be stored and the storage process
	Owner	Owner information (user name, title, etc)
	Account Type	Describes access level provided to owner
	Authorization	Authorization level necessary and the authorization process followed to gain access
Encryption	Entry Point	Determines if data are at rest or in transit
	Algorithms	List of potential encryption algorithms and processes to apply
	Application Method	Full disk, file system or database encryption and the steps that follow
Key Management	Storage	Separate storage space and maintenance to upkeep key storage
	Policy Management	Rules, regulations around key policies and access
	Authentication	Process by which user is authenticated to access keys and additional encryption information
	Key Transmission	Processes available for use in transmission of key information

Table 3.2: Data storage metadata for UC2

An example of these terms (in addition to some encryption-specific metadata) put into action is below. This example shows how metadata would be applied to the encryption process based off of current UC2 recommendations.

- Entry Point - determines if data are at rest or in transit
- Encryption Type - the cipher to use for data encryption, valid entries are AES, RSA, among others
- Encryption Key Length - the key length to use with the cipher in the CipherType parameter
- Application Method - full disk, file system or database encryption and the steps that follow

```
{
    "entry point": "in transit",
    "encryption type": "AES",
    "encryption key length": 256,
    "application method": "full"
}
```

An additional example relevant to UC2 surrounds key management. Key management is a critical component of storage and eventual access to data. Strong metadata practices are critical for key management in UC2. The below is an example of metadata for a specific key.

- Key Information - background information necessary to describe key
- Key Type - the type of key present
- Key Created Date - when the key was created, valid entries are in a MM-DD-YYYY string format.
- Key Owner - who owns the particular key
- Key Expiration - the expiration date of the key, valid entries are in a MM-DD-YYYY string format.
- Key Algorithm - algorithm the key is utilized in
- Sharing - determines if the key can be shared or not, on a scale of 0 (not sharable) to 1 (available to all)
- Audit Type - the audit performed or allowed on the key and its corresponding data
- Storage Location - where the key is stored
- Authentication - the level of authentication necessary to access the key

```
{
    "key information" {
        "key type": "HSM",
        "key created date": "01-01-2020",
        "key owner": "admin",
        "key expiration": "12-31-2020",
        "key algorithm": "RSA07",
    }
    "sharing": 0.2,
    "audit type": "admin only",
    "storage location": table317,
    "authentication": "text_passcode"
}
```

3.6 Data governance metadata for the eCommerce use case (UC3)

Anonymization, or transformations in general, can be applied at one of the following stages:

- at the *ingestion* phase into the data lake (see Section 2.3),
- in the *storage* phase (as considered here), and
- during the *analytics* phase (see Section 4.3).

Anonymization with differential privacy can be implemented in the following three models: the *local*, *central* or *hybrid privacy model* (the latter uses cryptographic protocols, e.g., secure computation), which are described in Section 2.3 of D2.1 (“Requirements from the Use Cases”). These models map to the different phases as described next.

In the local model, without a trusted third party, the data are already anonymized at the source. In our case, this happens at *ingestion* into the data lake and the necessary metadata are detailed in Section 2.3. In the central and hybrid model, sanitization is performed during the *analytics phase* w.r.t. to a certain analytical function.

One can consider the *storage phase*, e.g., when data are transferred from data lake to data market, to be implementable in either the local or central (hybrid) model. As mentioned before, the sanitization of data in the local model with the required metadata is detailed in Section 2.3. The central (and hybrid) model sanitization requires the same information as the local model sanitization, however, additionally a type of analytics can be given:

- `analyticsType`: pre-defined, supported analytical functions (e.g., sum, mean) with pre-set privacy parameters to reduce complexity for end users.

We give more detail for the central and hybrid models and their sanitization metadata in Section 4.3, as this sanitization is mainly used in the context of an analytics phase as detailed in Chapter 4.

4. Data Sharing and Analytics

In this chapter, semantics needed for the processing within the data market are explored. So far, the other metadata were ingested and processed in the data lake in order to prepare the data market. This means that metadata so far were available to help the transformations, sanitization and wrapping of data before they are put into the data market. Some of the metadata that were acquired during the ingestion or added as additional information to the data lake will still be present and can be used in the algorithmic control of what customers can see, process or extract from the data market.

Consequently, the current chapter only concerns the specific metadata that need to be present to support the operations *within* the data market and disregards all metadata that are only needed on the way of the data into the data market. As for the previous chapters, the metadata considerations are listed by use case. Duplicates are possible and not discouraged. The optimization will happen at the implementation phase when parts of similar algorithms can be re-used in more than one use case.

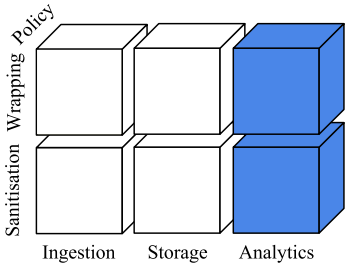

Figure 4.1: The MOSAICrOWN structure (analytics)

In this chapter we consider the metadata required to satisfy the requirements with regard to the MOSAICrOWN data life-cycle stages highlighted in blue in Figure 4.1.

4.1 Sharing and analytics metadata for ICV use case(UC1)

Metadata for the ICV use case are using and combining several vocabularies from the automotive industry and from privacy research and standardization. Table 4.1 reports the metadata related to data sharing and analytics. The categories and metadata are based on [Dat19] and [DPV19]. Note that there are duplicate entries in the sharing and analytics metadata and the governance in the data lake metadata. Consider the situation where a data owner only allows data to be stored in the data lake for a long period of time and only permits data to be used by a customer of the data market for a short period of time. In this situation, it is desirable to have a data retention period for data storage and a separate data retention period for sharing and analytics.

Category	Metadata	Description
Processing	Analyze	To study or examine the data in detail
	Make Available	To transform or publish data to be used
	Restrict	To apply a restriction on the processing of specific records
	Retrieve	To retrieve data, often in an automated manner
	Structure	To arrange data according to a structure
	Transfer	To move data from one place to another
	Transmit	To send out data
	Use	To use data
Technical and Organizational Measures	Access Control Method	Methods which restrict access to a place or resource
	Anonymization	Process by which personal data are irreversibly altered
	Authentication Protocols	Protocols involving validation of identity
	Authorization Procedure	Procedures for determining permission or authority
	Contract	Contractual terms governing data handling
	De Identification	Process by identifiable personal data (PII) are converted to un-identifiable personal data
	Encryption In Rest	Encryption of data when being stored
	Encryption In Transfer	Encryption of data in transit
	Legal Agreement	Legally binding agreement
	NDA	Non-disclosure Agreements
	Organizational Measure	Measures required/followed when processing data
	Privacy By Default	Practices regarding selecting appropriate data protection and privacy measures as the 'default' in an activity or service
	Privacy By Design	Practices regarding incorporating data protection and privacy in the design of information and services
	Storage Deletion	Defines how secure deletion is guaranteed
	Storage Duration	A duration or temporal entity denoting limitation on storage of personal data
	Storage Location	Defines a location or geospatial scope, where the data are (physically) stored
	Storage Restoration	Data restoration/backup mechanisms that guarantee that data are preserved
	Storage Restriction	Restrictions required or followed regarding storage of data
	Technical Measure	Technical measures required/followed when processing data of the declared category

Table 4.1: Metadata for sharing and analytics for UC1

An example of some of these terms is below. Note, these metadata are metadata related to

shared data, i.e., these metadata are related to the data when they are shared.

- Analyze - permission to analyze the data referenced
- Transfer - permission to transfer the data referenced
- Encryption In Transfer - method for encrypting data in transit

```
"ev-dpv-meta-data": {
    "veh_name": {
        "dpv:Analyze"  : "true",
        "dpv:Transfer" : "true",
        "dpv:EncryptionInTransfer" : "ECDHE-RSA-AES128-SHA256",
    }
}
```

4.2 Sharing and analytics metadata for the financial use case (UC2)

UC2 is using anonymization and encryption to make data marketable, but also allows for controlled sharing of data. The metadata used within the data market serve those purposes. They facilitate very granular data usage controls that also allow the provision of more details in case the necessary rights are acquired.

Category	Metadata	Description
Processing	Analyze	To study or examine the data in detail
	Make Available	To transform or publish data to be used
	Restrict	To apply a restriction on the processing of specific records
	Retrieve	To retrieve data, often in an automated manner
	Structure	To arrange data according to a structure
	Transfer	To move data from one place to another
	Transmit	To send out data
	Use	To use data
Technical and Organizational Measures	Access Control Method	Methods which restrict access to a place or resource
	Anonymization	Process by which personal data are irreversibly altered
	Authentication Protocols	Protocols involving validation of identity
	Authorization Procedure	Procedures for determining permission or authority
	Contract	Contractual terms governing data handling
	De Identification	Process by identifiable personal data (PII) is converted to un-identifiable personal data
	Encryption In Rest	Encryption of data when being stored
	Encryption In Transfer	Encryption of data in transit
	Legal Agreement	Legally binding agreement
	NDA	Non-disclosure Agreements

	Organizational Measure	Measures required/followed when processing data
	Privacy By Default	Practices regarding selecting appropriate data protection and privacy measures as the 'default' in an activity or service
	Privacy By Design	Practices regarding incorporating data protection and privacy in the design of information and services
	Storage Deletion	Defines how secure deletion is guaranteed
	Storage Duration	A duration or temporal entity denoting limitation on storage of personal data
	Storage Location	Defines a location or geospatial scope, where the data are (physically) stored
	Storage Restoration	Data restoration/backup mechanisms that guarantee that data are preserved
	Storage Restriction	Restrictions required or followed regarding storage of data
	Technical Measure	Technical measures required/followed when processing data of the declared category
Licenses and Data Usage	License Agreement	General terms of license signed with financial institution
	License Tier	Tier of financial institution as compared to level of security, involvement, etc. agreed to
	Geography	Geographical information for client and limitations on usage/sharing based on geography
	Application	Application(s) included in license along with previously inputted information from past engagement
	Special Instances	Any extraneous information pertinent to client from a license/DU perspective
	General Rules	Rules applicable to all clients, regardless of tier, with regards to data privacy

Table 4.2: Metadata for the data market for UC2

While the information detailed in the table is critical for ensuring overall sharing of data, one point that should be explained further is "permissible purpose". Metadata on the "permissible purpose" process will be critical for UC2, especially as the climate around data opt-in and usability gains traction throughout the world. The example below highlights the important steps needed to ensure cardholders have complete control around the usability and overall accessibility of their personal data.

- Identifier - used to identify the specific customer's data
- Data Type - the data type in question
- Intention of Use - list of potential use cases for data
- Permission Level - ability given from the identifier on whether the data can be used
- Conditions - any conditions specified from the user regarding use of data

- Disclosure - specifications on needs to disclose when data are used, these will be based off of the conditions specified
- Exceptions - any exceptions to when the conditions can be overturned and subsequent actions to alert data owner, these will be based off of the conditions specified

```
{
    "identifier": "2156",
    "data type": "transaction level data",
    "intention of use": "<useType>",
    "permission level": "use with conditions",
    "conditions": "<listofConditions>",
    "disclosure": "<disclosureConditions>",
    "exceptions": "<exceptionsConditions>"
}
```

4.3 Sharing and analytics metadata for eCommerce (UC3)

UC3 describes a Cloud-based data market for privacy-preserving consumer analytics. Anonymization is performed in the local, central or hybrid privacy model of differential privacy (see Section 2.3 in D2.1 ("Requirements from the Use Cases") for details). In the central and hybrid model, anonymization is applied in the *analytics* phase during evaluation of the data analytics query. Local differential privacy, on the other hand is used during data *ingestion* phase and before data *storage*, e.g., via addition of carefully selected noise at the data source before ingestion.

Sharing and analytics requirements

We recall the requirements for UC3 (Section 2.3 in D2.1) relevant for sharing and analytics:

- **Sanitization for local functions.** Data owners define anonymization at time of ingestion into the Cloud platform for sharing purposes.

  REQ-UC3-SL3: Sharing the result of the anonymization with data consumers (possibly data owners at the same time) through the cloud platform shall be possible.

- **Sanitization for central aggregation functions.** Data owners define anonymization for analytics on the Cloud platform.

  REQ-UC3-SC1: Anonymization parameters should be chosen by the data owner who executes the central data analytics function. The strength of the sanitization parameters is at the discretion of the data owner. This can be achieved by exposing aggregation functions through a fine-grained API to data owners.

  REQ-UC3-SC3: Collecting inputs from multiple data owners and evaluation of the aggregation function should be optionally possible via secure computation.

Sharing and analytics metadata

To fulfill these sanitization requirements, we need the following metadata:

- `dataSrc`: the identifier of the dataset that should be shared.

If the dataset was already anonymized at ingestion (local sanitization), no additional information is required. Otherwise, metadata that describe the central sanitization function are required, i.e., `method`, `methodConfig` consisting of `configParams`, `targetColumn`, and, optionally, `columnTypes`. See sanitization requirements in Section 2.3 for details.

Additionally, central sanitization is performed during the analytics phase w.r.t. to a certain analytical function, which the users need to select:

- `analyticsType`: a selection of preset types of analytics (e.g., sum, average) which is computed and sanitized.

The optional usage of secure computation, as in the hybrid privacy model, is enabled by selecting an appropriate analytics type (e.g., secure median).

5. Conclusions

This deliverable introduces the preliminary metadata model identified by the three industry use cases of MOSAICrOWN.

Via a process of requirements analysis and mapping the requirements to the data life cycle within MOSAICrOWN, it was possible to identify the following three categories of metadata required to define a model and language capable of supporting the data handling. From deriving these categories we conclude the following:

- *Data ingestion.* This category is concerned with the ingestion of data and metadata into the data market. The data ingestion phase of the data life cycle is the phase requiring the most metadata and therefore this category contains the largest number of metadata.

  The most significant amount of metadata required by MOSAICrOWN is at the ingestion stage. UC1 provides the majority of the metadata and this was primarily derived from FIWARE and W3C vocabularies and covered the topics of privacy, provenance, quality, GDPR, and encryption. UC2 introduces financial ontologies and vocabularies which will be used for ingesting the financial data. Additionally, UC3 introduces confidentiality, access control, and sanitization metadata. Finally, this category required the consideration of an ingestion API and the concept of semantification and policy ingestion.

- *Data governance in the data lake.* This category considers the threat model in the data lake, and the process which can be used in order to identify a process to mitigate the threats to the data lake. Furthermore, this category required the introduction of languages and metadata which can be used for access control and usage control within the data lake.

  From the perspective of the data lake, UC1 introduces the metadata required to govern the data within the metadata lake. These metadata encompass metadata such as the data controller, how long the data lake should be allowed to store the data, what geographies that are permitted to store the data, etc. UC2 requires investigation into the metadata required for data wrapping within the data lake such as the algorithms to be used, information regarding the key management and further topics. And finally, UC3 requires metadata relevant to how to sanitize the data in the data lake such as which stage and what privacy parameters to use.

- *Data sharing and analytics.* The final category of metadata covers the topic of data sharing and analytics from the perspective of the industrial use cases. Noteworthy in this chapter was how metadata required in the data governance category is also required in the data sharing and analytics category in different contexts.

  UC1 and UC2 share similar metadata however UC3 introduces other types of metadata required for sanitization. As can be expected, sanitization has significant relevance to data sharing and data analytics and as a result UC3 contributes more to this category than UC1 and UC2.

From the metadata identified from this deliverable it will be possible to define a policy model and language which in turn will be used by the policy management task.

Bibliography

[ABN+09] Claudio A Ardagna, Laurent Bussard, Gregory Neven, E Pedrini, S Paraboschi, F Preiss, P Samarati, S Trabelsi, M Verdicchio, and Sabrina De Capitani di Vimercati. Primelife policy language. In *Proc. of the W3C Workshop on Access Control Application Scenarios*, Luxembourg, November 2009.

[AI16] Riccardo Albertoni and Antoine Isaac. Data on the web best practices: Data quality vocabulary. https://www.w3.org/TR/vocab-dqv/, 2016. (Accessed on 02/28/2020).

[ASV+17] Sören Auer, Simon Scerri, Aad Versteden, Erika Pauwels, Angelos Charalambidis, Stasinos Konstantopoulos, Jens Lehmann, Hajira Jabeen, Ivan Ermilov, Gezim Sejdiu, Andreas Ikonomopoulos, Spyros Andronopoulos, Mandy Vlachogiannis, Charalambos Pappas, Athanasios Davettas, Iraklis A. Klampanos, Efstathios Grigoropoulos, Vangelis Karkaletsis, Victor de Boer, Ronald Siebes, Mohamed Nadjib Mami, Sergio Albani, Michele Lazzarini, Paulo Nunes, Emanuele Angiuli, Nikiforos Pittaras, George Giannakopoulos, Giorgos Argyriou, George Stamoulis, George Papadakis, Manolis Koubarakis, Pythagoras Karampiperis, Axel-Cyrille Ngonga Ngomo, and Maria-Esther Vidal. *The BigDataEurope Platform - Supporting the Variety Dimension of Big Data*, pages 41–59. Springer International Publishing, Cham, 2017.

[BHBL11] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[BKdM12] Diederik Bruggink, Pierre Karsten, and Carlo de Meijer. The european cards environment and iso 20022. *Journal of Payments Strategy & Systems*, 6(1):80–99, 2012.

[BW19] Daniel Bernau and Rigo Wenning, editors. *Deliverable D2.1: Requirements from the Use Cases*, Mosaicrown Deliverable, Dec 2019.

[CKPM05] Scott Cantor, John Kemp, Rob Philpott, and Eve Maler. Security assertion markup language (SAML) v2.0. Technical report, March 2005.

[Cou18] Cork City Council. Electric vehicle charge points - datasets - data.gov.ie. https://data.gov.ie/dataset/ev-charge-points, Jul 2018. (Accessed on 02/06/2020).

[Dat19] Data privacy vocabulary v0.1. https://www.w3.org/ns/dpv/, Nov 2019. (Accessed on 02/04/2020).

[DPV19] DPVCG GDPR legal basis vocabulary. https://www.w3.org/ns/dpv-gdpr, Jul 2019. (Accessed on 02/04/2020).

[DS05] Martin Duerst and Michel Suignard. Internationalized resource identifiers (IRIs). Technical Report 3987, January 2005.

[Eur16] European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), May 2016.

[Fin17] Financial industry business ontology, December 2017. Version 1.2, https://www.omg.org/spec/EDMC-FIBO/FND/1.2/PDF.

[GHM+14] Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, and Zhe Wang. HermiT: an OWL 2 reasoner. *Journal of Automated Reasoning*, 53(3):245–269, 2014.

[Har12] Dick Hardt. The OAuth 2.0 authorization framework. RFC 6749, October 2012.

[ISM+17] Renato Iannella, Michael Steidl, Mo McRoberts, Stuart Myles, James Birmingham, and Víctor Rodríguez-Doncel. Odrl vocabulary & expression. W3C Working Draft, available at https://www.w3.org/TR/2017/WD-odrl-vocab-20170223/, W3C, 2017.

[JG13] Andrew Jacobs and Marc Gratacos. Applying FpML. *Handbook on Systemic Risk*, page 66, 2013.

[KFD+18] Sabrina Kirrane, Javier D Fernández, Wouter Dullaert, Uros Milosevic, Axel Polleres, Piero A Bonatti, Rigo Wenning, Olha Drozd, and Philip Raschke. A scalable consent, transparency and compliance architecture. In *European Semantic Web Conference*, pages 131–136. Springer, 2018.

[LAC+19] Wonsuk Lee, Qing An, Adam Crofts, Kevin Gavigan, Justin (Jong Seon) Park, and Kevron Rees. Vehicle data. Technical report, September 2019.

[Liu13] Chunhui Liu. XBRL: a new global paradigm for business financial reporting. *Journal of Global Information Management (JGIM)*, 21(3):60–80, 2013.

[Mos05] Tim Moses. extensible access control markup language (xacml) v2.0. Technical report, 2005.

[MSAV16] Mohamed Nadjib Mami, Simon Scerri, Sören Auer, and Maria-Esther Vidal. Towards semantification of big data technology. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 376–390. Springer, 2016.

[OAS09] OASIS. extensible access control markup language (XACML) version 3.0, April 2009. http://www.oasis-open.org/committees/document.php?document_id=32425.

[PP19] Harshvardhan J. Pandit and Axel Polleres. Data privacy vocabulary v0.1, November 2019. https://dpvcg.github.io/dpv.

[Pro19a] FIWARE Data Models Project. Evchargingstation - fiware-datamodels. https://github.com/smart-data-models/dataModel.Transportation, Jun 2019. (Accessed on 02/04/2020).

[Pro19b] FIWARE Data Models Project. Github - smart-data-models/datamodel.transportation: Transportation data model. https://github.com/smart-data-models/dataModel.Transportation, Jun 2019. (Accessed on 02/28/2020).

[PTKH15] Rufus Pollock, Jeni Tennison, Gregg Kellogg, and Ivan Herman. Metadata vocabulary for tabular data. Technical report, Dec 2015. https://www.w3.org/TR/2015/REC-tabular-metadata-20151217.

[SDM19] Smart-Data-Models. Transportation harmonized data models, Oct 2019. https://github.com/smart-data-models/dataModel.Transportation.

[SLK+19] Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, Pierre-Antoine Champin, and Niklas Lindström. A json-based serialization for linked data. Technical report, Dec 2019. https://www.w3.org/TR/json-ld11.

[Veh20] Vehicle - fiware-datamodels, Feb 2020. https://fiware-datamodels.readthedocs.io/.

[W3C12] R2RML: RDB to RDF mapping language. Technical report, Sep 2012. https://www.w3.org/TR/r2rml.

[W3C13] W3C. PROV-overview, April 2013. https://www.w3.org/TR/prov-overview/.

[Wat17] Mark Watson. Web cryptography API, January 2017. https://www.w3.org/TR/WebCryptoAPI/.

[WJ15] Kim Wuyts and Wouter Joosen. LINDDUN privacy threat modeling: a tutorial. CW Reports, 2015. https://lirias.kuleuven.be/retrieve/331950.